Test Development

Introduction

The following information was developed by Air Force Occupational Measurement Squadron's Occupational Test Development Flight, 1550 5TH Street East, Randolph AFB TX 78150 4449.

Good test items are hard to come by. Many say that this is because writing good test items is "an art, not a science." To many examinees, this must indeed appear to be the case, because the world abounds with lackluster, artless tests. More often than not, however, the test-as-art point of view is just an excuse for a failure to take time and follow some basic rules. The fundamental rule is to keep in mind the importance of the test itself. The written test is literally the culmination of the training development effort and, as such, should be given the attention it deserves. Writing good test items is demanding and requires painstaking attention to detail. **Remember, it's not enough to write a test item so that it can be understood: you must write it so that it cannot be misunderstood.**

This handout will take you through the test-writing process, with an outline of procedures and advice that we think you will find helpful. The first section is a discussion of some basic rules to use when writing test items from scratch. The second section describes how statistical analysis can be used to fine-tune existing test items. The fnal section of the handout is a list of generic terms that you may find useful in developing test items.

Development of Test Items

The tests you develop must be valid and reliable in order to function as objective measures of learning. The more valid a test is, the better it measures what it was designed to measure, *e.g.,* a mechanics test should measure important knowledge in mechanics…not reading skills. The more reliable a test is, the better it provides consistent and predictable test results, *i.e.,* an individual should achieve the same score each time the same test is administered. This section discusses how test items should be developed to meet the basic prerequisites for validity and reliability. The following rules have been established by Air Force psychologists who have determined that the use of a standardized style and format of test items reduces item ambiguity and helps to increase overall test validity and reliability.

Rules for Item Writing

1. The item should focus on an important on-the-job situation. Ideally, items should focus on important problem situations that actually occur on the job and that require the examinee to know the correct answer to make a decision on the course of action to take, the tool to use, or the person to consult. The information m the item should be important for the examinee to know to perform the job adequately. The question should involve an aspect of knowledge that is relevant, realistic, and practical and

should test a student's ability to apply the knowledge rather than the ability to remember some rule without ever having the need to apply it. Just because some information appears in the course materials does not mean that you should necessarily honor it with a place on your test!

2. The stem should have a central problem that is clearly defined. This means it should contain enough information and be stated in such a way that the examinee knows exactly what's being asked. The following item has no central problem:

   Which of the following statements best describes the side-looking function?
   A) It is a part of all fighter avionics systems
   B) It is a part of all bomber avionics systems
   C) It should be activated by the radar switch during the boost phase
   D) It should be activated by the radar switch during the climb phase

   A simple test to see whether or not the problem is clearly posed is to read the stem alone, covering up the choices. If you have to read all the choices before being able to answer the question, the stem is not clear. If the central problem in the sample question really is "What switch should be activated and during what flight phase?" (as indicated by the answer), then the item should be rewritten to clarify that:

   What flight control switch should be activated to achieve the side-looking function, and during what phase of flight should the switch be activated?
   A) Radar switch during climb phase
   B) Radar switch; during boost phase
   C) Search-scan switch; during climb phase
   D) Search-scan switch; during boost phase

   The item now has a central problem because the problem posed is clearly stated and there is no question as to what information is being sought. Just by reading the stem you can answer the question without having to read all of the choices. The choices fulfill other item writing requirements because they are parallel to each other in content and style of language and are all responsive to the generic terms "flight control switch" and phase of flight. " (Refer to Rules 4, 8, and 9.)

3. Just the right amount of information should be used in the stem to state the central problem. Extra information in the stem increases the reading time, confuses the examinee, and may ruin an otherwise good item. The following stem has too much information:

   The B-1 is a supersonic bomber designed to replace later model B-52's. What type of engine does it have?

   The first sentence is wasted because it is not necessary to answer the question. The stem of this item could be rewritten in the following way so that it contains only the essential information required to answer the question:

   What type of engine does the B-1 have?

In the same way, an insufficient amount of information in the stem may not only confuse examinees but may result in an item that does not have a technically correct answer or may be interpreted to have more than one correct answer because of its ambiguity.

4. A generic term should normally be used in the stem. A generic term is a word or phrase that refers to all four choices and normally appears **after** the interrogative word (e.g., "what" or "which of the following), and **before** the verb in the stem. The purpose of wing a generic term is to make the item more understandable to the examinees. The following item lacks a generic term:

   What supplies fuel to an internal combustion engine?
   A) Oil pump
   B) Generator
   C) Alternator
   D) Carburetor

   The stem would better communicate with a generic term included:

   What component supplies fuel to an internal combustion engine?

   The word "component" is the most appropriate generic term to we in the stem because all of the choices can be considered to be components.

   The last section of this handout is a list of generic terms you may find helpful.

5. There must be only ONE correct answer for each item. You must make sure that the distracters you we are not correct, that the correct answer listed as a choice is the only correct answer for the situation presented in the stem, and that the answer is not contradicted by any other directive. You should spend a sufficient amount of time researching the reference material to ensure that the distracters are not serving as correct answers. Often, test items have to be deleted from scoring because of multiple correct answers. This type of error can be prevented if item writers spend a few additional minutes on each question to ensure that the answer is fully supported by the reference and that the distracters are not correct. It is also important to note that you never include correct information in distracters simply because the information cannot be found in the available reference material.

6. Each item must be supported by a specific and complete reference. The reference used to write the item and extract the correct answer must be entered accurately and completely on the Item Record Card. The reference entered must be complete enough so that anyone later reviewing the item can locate the correct answer from the reference source you provided. Having a cited and valid reference for each item in a test is one way of ensuring that the test is fair to everyone taking it.

7. Distracters must be plausible but not correct. Each distracter must be a real thing or a believable, reasonable, and possible idea or action. Distracters should look like

possible correct answers to examinees who, although they have been in the specialty for several years, may not know the answer. The use of plausible distracters helps to "distract" examinees from guessing at the correct answer. Ideally, distracters should consist of common mistakes made on the job. The following item has a distracter that is not plausible:

Which of the following conditions could cause an automobile engine to overheat?
A) Faulty fuel pump
B) Loose door handle
C) Faulty water pump
D) Cracked distributor cap

Distracter B is obviously incorrect, even to someone with very little knowledge of automobile engines. The examinee who does not know the answer has only three choices to guess from instead of four; thus, the chance of guessing the correct answer is increased.

8. Choices should be responsive to the stem. Choices that are not responsive to the stem are those that do not logically answer the question asked. For example:

Why is silicon more suitable as a filament material for vacuum tubes?
A) High melting point
B) High work function
C) Low melting point
D) Low vaporization point

The choices do not answer the question of "why" and the stem is incomplete - silicon is more suitable than what? The following item offers a solution to the problem:

What characteristic of silicon makes it more suitable than elastic as a filament material for vacuum tubes? ~A) High melting point B) High work function C) Low melting point D) Low vaporization point

Items that lack responsiveness are usually items with stems that lack sufficient information, lack a central problem, and/or lack an appropriate generic term. Choices that are not similar in content and style of language are also not responsive to the stem.

9. Choices should all be similar (parallel! in content and style of language so that none of the choices will stand out as being different. Choices that stand out increase the chance of guessing the correct answer. The choices in the following item are not similar in content or style of language:

Which of the following tools should be used to anneal carbon steel?
A) A blowtorch
B) Cyanide solutions
C) Quenching in water

D) A circulating-air furnace

In addition to the choices being dissimilar, using the word "tool" as a generic term has made it easy to guess the correct answer since the correct answer is the only tool listed as a choice. (This is considered a clue. See Rule 12.) The distracters are not similar in content or style of language in that they are not all responsive to the generic term used. This item is an example of "mixing apples with oranges." As specified in Rule 4, the generic term used should refer to all the responses to make the item more understandable to the examinees. An item such as the one above would help examinees who do not know the answer to guess it correctly, not because of job knowledge but because the answer is the only logical response based on the generic term used. This type of item is of no value in a test of task knowledge. The item can be improved by using either one of the following approaches:

Which of the following tools should be used to anneal carbon steel?
A) Hammer
B) Blowtorch
C) Screwdriver
D) Welding torch

or

How should carbon steel be annealed?
A) By quenching the steel in water
B) By heating the steel with a blowtorch
C) By heating the steel in a circulating-air furnace
D) By applying cyanide solution to the steel after cooling

In the first approach, the distracters were changed to make them tools so that they could be responsive to the generic term used. In the second approach, a different question was asked, and all the choices were modified to make them all similar in content, style of language, and responsive to the question "How." The style of language used in the choices is similar in that the choices begin with the word "By" and are followed by an action verb, e.g., applying, heating.

10. Choices must be discrete; they must stand-alone and not be part of. or included in. another response. This is a prerequisite to ensuring that an item has only one correct answer. The following item shows the kind of problems encountered when inclusive choices exist because the stem is not clear:

At what age is an individual eligible for the draft?
A) 18
B) 19
C) 26
D) 36

The stem is not precise. It is not clear whether it refers to the age at which an individual first becomes eligible, the maximum age at which he is eligible if deferred, or any of the ages between these limits. Hence, B, C, and D could be considered correct answers. Two possible solutions:

If an individual receives a school deferment, at what age will he no longer be eligible for the draft?
A) 17
B) 18
C) 26
D) 36

At what age does an individual first become eligible for the draft?
A) 17
B) 18
C) 19
D) 21

Always be careful when referencing this type of item, particularly when asking for a minimum and a maximum. Ensure that the stem is fully qualified so that the question that is being asked is not misinterpreted. Often the wording can be tricky, and this can lead to multiple correct answers because of item ambiguity.

11. Choices should be approximately the same length to ensure that all choices are presented in the same general form and that one choice does not stand out. The correct answer may be given away if it stands out by being much longer and/or more general than the other choices. In the same way, a short, precise, correct answer with long complex distracters will stand out and may give the answer away. If necessary, two sets of paired choices, where the pairs are of slightly different length, may be used if they are plausible.

Acceptable:

A) _______
B) _______
C) ___________
D) ___________

A) ________
B) ___________
C) _______________
D) _______________

Not acceptable:

A) ______
B) ______
C) ______
D) ___________

A) _____
B) ________________
C) ________________
D) ________________

12. Do not provide clues to correct answers. One of the ways in which the correct answer can be clued is when a key word or phrase in the stem also appears in the correct answer and not in the other choices. This type of question should be avoided because it could cause a guessing examinee to get the correct answer unfairly. Also, whenever possible, avoid repeating a key word from the stem in only one distracter because this makes the distracter stand out. The following item clues the correct answer:

    Which of the following resources should be used by a dental clinic to determine proper routing of a request for referral?
    A) AFR 16W3, Medical Examination and Medical Standards
    B) AFR 161-23, Aerospace Medicine Consultant Service
    C) AFR 161-36, Medical Recommendation for Flying or Special Operational Duty
    D) AFR 162-1, Management and Administration of USAF Dental Activities

    The word "dental" in the stem also appears in the correct answer but in none of the distracters and, therefore, may give the answer away to guessers who are looking for a clue. One way to eliminate this problem would be by repeating the key word in one of the distracters. The question below has been corrected so that the key word in the stem does not clue the answer.

    Which of the following resources should be used by a dental clinic to determine proper routing of a request for referral?
    A) AFR 16043, Medical Examination and Medical Standards
    B) AFR 161-23, Aerospace Medicine Consultant Service
    C) AFR 162-1, Management and Administration of USAF Dental Activities
    D) AFR 162-7, US Air Force Dental Investigation Service

    Since distracter D now has the word "dental," the correct answer is no longer uniquely clued. The problem of providing clues to answers can also be corrected by eliminating the key word in the stem or the choice. (Another example of providing a clue to an answer is illustrated in Rule 9.)

13. Items that refer to forms or publications should include both the number and title. This rule is based on a fact that a student's mastery of specialty knowledge does not include the requirement to memorize form or publication numbers and titles. You should make use of AFIND 2, Numerical Index of Standard and Recurring Air Force Publications, and AFIND 9, Numerical Index of Departmental Forms, when initially drafting a new item or when re-referencing the current tests to ensure that the numbers and titles used are current and accurate. Better yet, check the form or publication itself, if it is available.

14. Negative items should be avoided. Most negative items only succeed in confusing the examinee taking the test because they reverse the examinee's normal way of thinking. Consider the following item:

    Which of the following items of equipment should not be issued to aircrew members?

A) Flashlight
B) B) First-aid kit
C) Oxygen mask
D) Hunting knife

Distracters A, C, and D are items that should be issued to aircrew members, while the correct answer is an item that is not issued because it is already available in the aircraft for all aircrew members to use. This item is confusing because it requires the examinee to think in a manner contrary to the manner of thinking on the job. As mentioned before, our attempt in writing test items should be to ask questions that require an examinee to know what course of action should be taken or how the problem should be solved, and distracters should be mistakes commonly made on the job. Negative items are sometimes appropriate to test dangerous or forbidden practices where knowledge of these exceptions is very important for reasons such as safety. Negative items are acceptable only under the following conditions: (1) it is the most direct method for extracting an examinee's knowledge of a dangerous or forbidden practice, and (2) a specific reference can be found which cites the practice as being dangerous or forbidden. When negative items are appropriate, words such as **not, never,** and **except** must be given special emphasis in the stem. The following example illustrates as acceptable negative item:

Which of the following drugs should never be administered to an individual with severe chest pains?
A) Morphine
B) Adrenalin
C) Propanolol
D) Nitroglycerine

Using the Hierarchy of Content Validity

To help you write questions with varying degrees of difficulty, Figure 3 relates proficiency levels as shown on an STS to the kinds of questions one can write. This table is designed to show the increase of proficiency, which is required, as a person advances in grade as well as skill level. In addition, it offers suggestions for question formats which might best test that degree of proficiency.

The following examples demonstrate the development of three different items (each for a different test level) from a single reference:

The Reference

> The purpose of the carburetor in an internal combustion engine is to convert fuel into a gaseous state by mixing it with air. This mixture is delivered to the engine on each intake stroke of a piston and ignited on the power stroke.

A-Level Item. This item calls for simple recall of an equipment part or component and does not require the examinee to know much about the operation of an internal combustion engine to answer the question correctly.

What component supplies fuel to an internal combustion engine?
A) Oil pump
B) Generator
C) Alternator
D) Carburetor

B-Level Item. This item requires the examinee to understand something about the procedure through which the carburetor performs its function.

What is the purpose of a carburetor?
A) To ensure that sufficient vacuum is maintained
B) To distribute the fuel-air mixture to the engine
C) To mix fuel with air and deliver the mixture to the engine
D) To extract air from the fuel and deliver the fuel to the engine

C-Level Item. This item requires the examinee to have an understanding of a more complex functional procedure and indirectly tests the examinee's understanding of the principle behind the operation of internal combustion engines.

How does a carburetor furnish a fuel charge to an engine?
A) By mixing fuel, air, and oil
B) By converting fuel to a gaseous fuel-air mixture
C) By injecting the liquid fuel directly into the engine
D) By injecting the fuel-air mixture directly into the engine

Figure 3. Hierarchy of Content Validity

STS/PII Level	Proficiency level required on the job	Sentence styles
1/a/A	➢ Can do simple parts of the task. ➢ Can name parts, tools, and simple facts about the task. ➢ Can identify basic facts and terms about the subject.	• What is the goal of . . .? • What are the parts of a . . .? • What is the correct way to . . .? • What type of . . . is used for . . .? • What component (some function) . . .? • What are the categories of . . .? • What tools should be used to . . .? • Where is (a piece of equipment) located on a …? • What quantity of transactions should be . . .?
2/b/B	➢ Can do most parts of the task. ➢ Can determine step-by-step procedures for doing the task. ➢ Can explain relationship of basic facts and state general principles about the subject.	• What is the purpose of . . .? • When should (some actin) be taken? • What should be the first step in determining . . .? • What is the relationship between . . . and . . .? • How should (some form) be annotated when . . .? • In what sequence should (some procedure) be accomplished? • What activity must be contacted to obtain approval before . . .? • Which of the following actions should be taken to correct (some malfunction)?
3/c/C	➢ Can do all parts of the task quickly and accurately. ➢ Can tell or show others to do the task. ➢ Can explain why and when the task must be done and why each step is needed. ➢ Can predict, identify, and resolve problems about the task. ➢ Can analyze facts and principles and draw conclusions about the subject.	• How should . . . be determined? • How does (some system operate? • How should (some malfunction) be corrected? • What activities are exempt from . . .? • Why should (some action be taken) when (some situation Occurs)? • What condition is indicated by . . . when a . . .? • What action should be taken if . . .? • If (some condition) occurs, what action should be taken, and why? • What factors should be considered when determining . . .? • What condition would cause (some system/component) to malfunction, and what corrective action should be taken? • What situations require the use of . . .? • Why should (some activities/unit) be contacted when . . .?

LIST OF GENERIC TERMS

abilities
actions
activities
adjustments
advantages
agencies
agents
aids
alternatives
amounts
appearances
applications
areas
arrangements
assemblies
assumptions
attempts
authorities
authorizations

bases
benefits

calculations
capabilities
categories
causes
changes
characteristics
checks
circumstances
classifications
combinations
complications
components
computations
concentrations

concerns
conclusions
conditions
considerations
constraints
containers
controls
corrections
criteria

damages
dangers
decisions
defects
deficiencies
devices
differences
difficulties
dimensions
directions
directives
disadvantages
discrepancies
diseases
disparities
disposition
documents

effects
elements
entries
equipment items
errors
essentials
estimations
events
examinations
exceptions

facilities
factors
failures
faults
features
forces
forms
formats
frequencies
functions
gains
groups

impairments
imperfections
implications
improvements
indications
individuals
influences
information
ingredients
instances
instructions
instruments
irregularities
items

levels
lists
locations

malfunctions
management
tools
markings
materials
means
measurements
mechanisms
metals
methods
modes
models
modifications

notations

objects
objectives
observations
occurrences
offices
operations
organizations

parts
patterns
personnel
points
policies
positions
possibilities
precautions
prerequisites
principles
problems
procedures
processes
products
provisions
publications
purposes

qualifications
qualities
quantities
ranges
rates
readings
reasons
recommendations
references
relationships
repairs
requirements
responses
results

schedules
sections
segments
sequences
services
settings
signs
situations
solutions
sources
specifications
stages
standards
steps
substances
subsystems
symbols
systems

tasks
techniques
theories
time periods
tools
transactions
transmissions
types

units
uses

values
variations
varieties

warnings
ways
weaknesses